M - 35

THE METHODS OF VERTEX DISCRIMINANT MULTICATEGORY ANALYSIS WITH BOOTSTRAP APPLICATION

Warnida Lena[1], I Made Sumertajaya, Bagus Sartono
Statistics Department FMIPA, Bogor Agricultural University, Indonesia
E-mail[1]: warnidalena6@gmail.com

Abstract

Underdeveloped district is the district that has less developed community and region compare with the other areas in the national scale base on economical category, society, human resources, infrastructure, financial capacity, accessibility, and regional characteristics. It is not easy in classification of underdeveloped district that involves many variables and the number of observations by multicategory cases. Sometimes the data used does not fill the double normal assumption and the group of variance-covariance matrix is not a homogeneous. Verteks Discriminant Analysis (VDA) is the method of the newest multicategories classification which can handle high-dimensional data. In this research, the significance testing of the function of vertex discriminant is done by using a bootstrap approach to determine the significantly variables. Through simultaneous confidence intervals of $\boldsymbol{T^2}$ **– Hotelling**, based on the results of analysis show that from the 27 predictor variables used, all significant variables have the real effect in determining the status of underdeveloped district. The accuracy of the classification that is obtained from the vertex discriminant function is about 94.44% for data training and 72.22% for data testing.
Keywords: underdeveloped district, VDA, bootstrap, simultaneous confidence interval

INTRODUCTION

Ministry of Rural Development (KPDT) states that the development of underdeveloped areas is a deliberate attempt by the government to change the area with various socio-economic problems and physical limitations into developed areas with the same quality of life or not far behind compared to other Indonesian regions. KPDT determines the status of regional lag based on six main criteria, namely economy, society, human resources, infrastructure, financial capacity, accessibility, and regional characteristics. Status of the lag will be used as the references of rural development.

Based on the guidelines of the Explanation of Determination of Underdeveloped Areas by KPDT, explained that technical calculation of underdeveloped area is done by multiplying the results of standardization data with the weights for each variable, then multiplicate ± 1 where the indicator to measure the level of ugliness become positive, then do the summation **.** The total is used as a benchmark of the determination of underdeveloped district status. However, the standardization process will generate positive and negative value. When doing the multiplication to ± 1, it will give the opportunity to produce the wrong conclusions.

The appropriate grouping of underdeveloped area based on the causing factor of its lag is a very important issue to support the strategies of the coping lag areas. For that, it takes the appropriate analytical methods to get more accurate and efficient results. The common classification method used in classifying an object is the discriminant analysis.

Discriminant analysis is a statistical method that is commonly used in classifying an object and allocates a new object into a group that has been defined previously. Discriminant function gives the value as close as possible to the objects in the same group and as far as

possible for the objects between groups (Rencher, 2002). There are several discriminant analysis which are currently developing include Fisher linear discriminant analysis (ADF), quadratic discriminant analysis, canonical discriminant analysis and the latest linear discriminant analysis namely vertex discriminant analysis (VDA). Nurmaleni (2014), conducts a study that compares two methods of linear discriminant ADF and ADV. In its application, the ADF can not be used for the classification when the X matrix is not full rank. It causes the singular variance-covariance matrix so that it has no inverse matrix. This condition will happen when many greater variables *p* rather than many observations *n*.

VDA is the latest multicategory classification method which is introduced by Lange and Wu (2008). VDA can classify the objects when *X* matrix is full rank and not full rank. The classification in VDA is done by minimizing the objective function that involves $\in$ -insensitive loss and quadratic penalty.

Bootstrap is one of the estimation technique of confidential interval of population parameter that is able to cope the deviated data from its assumptions and the data which does not have the distribution assumptions (Efron and Tibshirani, 1993).

This research will examine how to determine the variables that have a real influence in the formation of discriminant function of VDA multicategory method for the full rank of matrix X by using bootstrap resampling approach through $T^2 -$ Hotelling simultaneous confidential intervals.

RESEARCH METHOD

Method and Material

The data used in this research were 162 data underdeveloped districts in Indonesia, which was obtained from the Ministry of Rural Development which sourced from the PODES collection In 2011 and 2014, SUSENAS in 2012, SUSENAS in 2013 and Financial Statistics of Regency/City in 2012-2013 (sourced from Central Bureau of Statistics). While the variables used in this study can be seen in Table 2.

Table 2 Variables used in modeling of underdeveloped district

Variable(s)	Annotation
Y	Less underdeveloped districts (1) Underdeveloped districts (2) Very underdeveloped districts (3)
X_1	The percentage of poor citizen
X_2	The outcome of consumption per capita
X_3	Life expectancy
X_4	The average length of school
X_5	Literacy rate
X_6	The number of villages with the widest asphalt/concrete road surface types
X_7	The number of villages with the widest paved road surface types
X_8	The number of villages with the widest land road surface types
X_9	The number of villages with the widest other types of road surface
X_{10}	Percentage of households with electricity users
X_{11}	Percentage of households with telephone users
X_{12}	Percentage of households with clean water users
X_{13}	The number of villages which have no permanent building market
X_{14}	The number of health facilities per 1000 population
X_{15}	The number of physicians per 1000 population

X_{16}	The number of elementary and junior high schools per 1000 population
X_{17}	Financial capacity
X_{18}	The average distance from the village office to the district office
X_{19}	The distance from the village to the basic education services
X_{20}	The number of villages with access to health services > 5 km
X_{21}	Percentage of vulnerable villages of earthquakes
X_{22}	Percentage of vulnerable villages of landslides
X_{23}	Percentage of vulnerable village of flood
X_{24}	Percentage of vulnerable village of other disaster
X_{25}	Percentage of villages in protected areas
X_{26}	Critical land village
X_{27}	Percentage village of conflict in recent year

Some of the tools used in this study are R 3.0.3 with package VDA, Minitab 16 and MS Excel.

Methodology

The Formation of Verteks Discriminant Function

The steps of the data analysis as follows:

1. Separating the district data into strata based on the status of underdevelopment;
2. Doing resampling bootstrap in each strata with n data collecting as much as available data on each strata, $n_l = n_{l'}$, where $l = l' = 1,2,\dots,k$;
3. Combining the data from collecting each strata;
4. Standardize all predictors to have mean 0 and variance 1;
5. The formation of VDA function with following stages: (Lange and Wu 2008)
   a. Deciding the initial of iteration $\mathrm{m} = 0$ with $\boldsymbol{A}^{(0)} = 0$ and $\boldsymbol{b}^{(0)} = 0$;
   b. Determining the value of vertices from each group by the equation:

$$\mathbf{v}_j = \begin{cases} (k)^{-\frac{1}{2}}\mathbf{1} & \text{if } j=1 \\ \mathrm{c}\mathbf{1}+\mathrm{d}\mathbf{e}_{j-1} & \text{if } 2 \le j \le k+1 \end{cases} \qquad \mathrm{c} = -\frac{1+\sqrt{k+1}}{(k)^{\frac{3}{2}}} \text{ and } d = \sqrt{\frac{k+1}{k}}$$

   and defines $\mathbf{y}_i = \boldsymbol{v}_j$;

   c. Majoring the loss function $\mathrm{R}(\mathbf{A},\mathbf{b}) \le \frac{1}{n}\sum_{i=1}^{n} \mathrm{w}_i \left\|\mathbf{r}_i - \mathbf{s}_i\right\|^2 + \lambda \sum_{j=1}^{k} \left\|\mathbf{a}_j\right\|^2$ with ith current residual $\mathbf{r}_i^{(\mathrm{m})} = \mathbf{y}_i - \mathbf{A}^{(\mathrm{m})}\mathrm{x}_i - \mathbf{b}^{(\mathrm{m})}$ and case weights:

$$\mathrm{w}_i = \begin{cases} \dfrac{1}{2\|\mathbf{r}_i^{\mathrm{m}}\|}, & \text{for } \|\mathbf{r}_i^{\mathrm{m}}\| \ge 2\epsilon \\ \dfrac{1}{4(\epsilon - \|\mathbf{r}_i^{\mathrm{m}}\|)}, & \text{for } \|\mathbf{r}_i^{\mathrm{m}}\| \le \epsilon \\ \dfrac{1}{4(\|\mathbf{r}_i^{\mathrm{m}}\| - \epsilon)}, & \text{for } \epsilon \le \|\mathbf{r}_i^{\mathrm{m}}\| \ge 2\epsilon \end{cases}$$

$$\mathbf{s}_i = \begin{cases} \mathbf{0} & \text{for } \|\mathbf{r}_i^{\mathrm{m}}\| \ge 2\epsilon \\ \mathbf{r}_i^{\mathrm{m}} & \text{for } \|\mathbf{r}_i^{\mathrm{m}}\| \le \epsilon \\ \left(\dfrac{2\epsilon}{\|\mathbf{r}_i^{\mathrm{m}}\|} - 1\right)\mathbf{r}_i^{\mathrm{m}} & \text{for } \epsilon \le \|\mathbf{r}_i^{\mathrm{m}}\| \ge 2\epsilon \end{cases}$$

   d. Minimize the surrogate function by determining $\mathbf{A}^{(\mathrm{m}+1)}$ and $\mathbf{b}^{(\mathrm{m}+1)}$ that is obtained from solving k sets of linear equations;
   e. If $\left\|\mathbf{A}^{(\mathrm{m}+1)} - \mathbf{A}^{(\mathrm{m})}\right\| < \gamma$ and $\left|\mathrm{R}(\mathbf{A}^{(\mathrm{m}+1)}, \mathbf{b}^{(\mathrm{m}+1)}) - \mathrm{R}(\mathbf{A}^{(\mathrm{m})}, \mathbf{b}^{(\mathrm{m})})\right| < \gamma$ both hold for $\gamma 10^{-4}$, then stop;

f. Otherwise repeat steps c through e.

6. Doing the step 2 to 5 by resampling bootstrap as $B = 1000$ (Davison and Hinkley, 1997). It will be obtained vertex discriminant function as follows:

Resampling bootstrap $B = 1000$

$$Y_{1.1} = a_{1.1.1}X_1 + a_{2.1.1}X_2 + a_{3.1.1}X_3 + \cdots + a_{p.1.1}X_p + a_{0.1.1}$$
$$Y_{1.2} = a_{1.1.2}X_1 + a_{2.1.2}X_2 + a_{3.1.2}X_3 + \cdots + a_{p.1.2}X_p + a_{0.1.2}$$
$$\vdots$$
$$Y_{1.1000} = a_{1.1.1000}X_1 + a_{2.1.1000}X_2 + a_{3.1.1000}X_3 + \cdots + a_{p.1.1000}X_p + a_{0.1.1000}$$

$$Y_{2.1} = a_{1.2.1}X_1 + a_{2.2.1}X_2 + a_{3.2.1}X_3 + \cdots + a_{p.2.1}X_p + a_{0.2.1}$$
$$Y_{2.2} = a_{1.2.2}X_1 + a_{2.2.2}X_2 + a_{3.2.2}X_3 + \cdots + a_{p.2.2}X_p + a_{0.2.2}$$
$$\vdots$$
$$Y_{2.1000} = a_{1.2.1000}X_1 + a_{2.2.1000}X_2 + a_{3.2.1000}X_3 + \cdots + a_{p.2.1000}X_p + a_{0.2.1000}$$

$$Y_{k.1} = a_{1k.1}X_1 + a_{2.k.1}X_2 + a_{3.k.1}X_3 + \cdots + a_{p.k.1}X_p + a_{0.k.1}$$
$$Y_{k.2} = a_{1.k.2}X_1 + a_{2.k.2}X_2 + a_{3k.2}X_3 + \cdots + a_{p.k.2}X_p + a_{0.k.2}$$
$$\vdots$$
$$Y_{k.1000} = a_{1.k.1000}X_1 + a_{2.k.1000}X_2 + a_{3.k.1000}X_3 + \cdots + a_{p.k.1000}X_p + a_{0.k.1000}$$

The Significance Test of Vertex Discriminant Function

The significance test of VDA discriminant function method is done by determining the 95% simultaneous confidence intervals ($T^2 - Hotelling$) for each discriminant coefficients (Johnson dan Wichern 2007).

$$\bar{a}_i - \sqrt{\frac{(n-1)p}{(n-p)}F_{\alpha;p;n-p;}}\sqrt{\frac{s_{ii}}{n}} \leq \mu_{a_i} \leq \bar{a}_i + \sqrt{\frac{(n-1)p}{(n-p)}F_{\alpha;p;n-p;}}\sqrt{\frac{s_{ii}}{n}} \quad (1)$$

where i = 1, ..., p which all have confidence coefficient of 1-α. The simultaneous confidence intervals for the coefficients of variable discrimination which contain a value of zero indicate that the variables do not give the real effects in the vertex discriminant function.

New Object Classification

The classification of new objects in the VDA method is predicted by (Lange and Wu, 2010):

$$\hat{y} = \text{argmin}_{j=1,..,k+1} \left\| \boldsymbol{v}_j - \hat{\mathbf{A}}x_i - \hat{\mathbf{b}} \right\| \quad (2)$$

The Size of Error Classification in Discriminant Analysis

Apparent error rate (APER) is one of methods to evaluate the discriminant analysis in the classification. According to Johnson and Wichern (2002) APER value is many error percentages in grouping by the function of classification.

$$\text{APER} = \frac{\text{the amount of wrong objects in the classification}}{\text{the amount of objects}}$$

RESULTS AND DISCUSSION

The data used for the formation of discriminant functions were taken from 162 underdeveloped districts in Indonesia. The data is divided into 144 districts as training data and 18 districts as testing data. The description of training data for each variable in each group can be seen in Table 3.

Tabel 3 Description of underdeveloped districts data with each variable in each group

Variables	Less underdeveloped districts		Underdeveloped districts		Very underdeveloped districts	
	Mean	Stdev	Mean	St.dev	Mean	Stdev
X_1	15,62	6,03	24,66	8,73	25,79	11,88
X_2	622,44	14,06	612,71	17,22	593,04	39,74
X_3	67,12	2,41	65,85	2,31	66,81	2,40
X_4	7,65	0,78	7,06	1,14	5,77	1,66
X_5	93,94	4,48	87,78	7,79	69,74	26,25
X_6	100,67	72,34	58,65	42,90	38,33	39,92
X_7	42,72	49,00	47,60	39,80	25,04	25,25
X_8	25,04	36,55	46,40	60,53	80,68	137,64
X_9	1,93	4,08	3,35	6,53	1,76	2,88
X_{10}	75,84	23,75	55,52	29,63	45,27	34,52
X_{11}	3,08	2,54	2,37	2,42	2,08	2,61
X_{12}	53,66	24,21	50,85	29,12	32,01	26,71
X_{13}	15,93	22,78	13,00	12,02	11,24	20,68
X_{14}	0,42	0,48	0,47	0,40	0,27	0,38
X_{15}	0,06	0,06	0,06	0,06	0,05	0,06
X_{16}	0,38	0,33	0,29	0,23	0,21	0,25
X_{17}	2056043,32	17875251,15	509498,58	940500,77	348670,06	160725,46
X_{18}	58,65	44,43	72,97	47,67	121,35	170,47
X_{19}	14,81	13,02	22,61	13,98	28,13	25,23
X_{20}	10,41	8,60	13,53	9,29	11,57	11,88
X_{21}	4,40	8,95	4,37	10,14	7,24	12,93
X_{22}	7,95	9,82	12,40	13,77	9,04	10,45
X_{23}	24,70	17,85	15,24	12,89	10,50	10,21
X_{24}	26,60	23,25	30,47	28,96	24,70	38,44
X_{25}	4,44	8,33	6,74	10,34	11,70	16,31
X_{26}	19,61	17,16	23,06	21,62	36,45	33,25
X_{27}	4,23	5,09	4,22	3,25	6,11	5,87
N	103		20		21	

Table 3 explains that the average percentage of the poor citizen (X_1) in the group of less underdeveloped district status is about 15.6 with standard deviation of 6.03. Then, in the group of underdeveloped district status, the average percentage of the poor is about 24.66 with standard deviation of 8.73. In the group of highly underdeveloped district status, the average percentage of the poor is about 25.79 with standard deviation of 11.88. The percentage indicator of the poor citizen (X_1) is an indicator that measures the badness. It can be seen that the higher value of the average percentage of the poor citizen (X_1) indicates the status of underdeveloped district is getting worse.

It also shows that for the indicator of the expenditure of consumption per capita (X_2) in the group of less underdeveloped district status is about 622.44 with standard deviation of 14.06, in the group of underdeveloped district status is about 612.17 with

standard deviation of 17.22 and in the group of highly underdeveloped district status is about 593.04 with standard deviation of 39.74. The indicator of the expenditure of consumption per capita (X_2) is an indicator that measures the goodness. The greater average value of the indicator of the expenditure of consumption per capita (X_2) indicates the status of underdeveloped districts is getting better. And then, for the indicator that measures the goodness, the higher value of indicator of district status tends to be better and the higher value of indicator that measures the badness of district status tends to be worse. Although it is not happened at all indicators.

In training data, districts with the status of less underdeveloped are 103 districts, districts with the status of underdeveloped are 20 districts and districts with the status of highly underdeveloped are 21 districts. In testing data, districts with the status of less underdeveloped are 12 districts, districts with the status of underdeveloped are 3 districts and districts with the status of highly underdeveloped are 3 districts. Districts with the status of less underdeveloped most nearly spread throughout the province in Indonesia. While the district with the status of underdeveloped and highly underdeveloped more spread in eastern Indonesia.

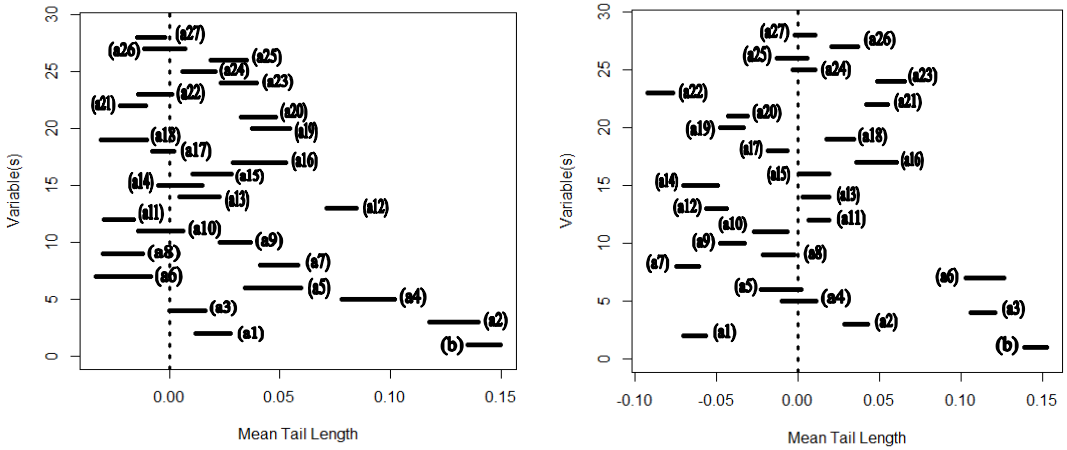
The Formation of Verteks Discriminant Function

By 1000 *bootstrap* replicates, VDA method forms each 1000 vertex function Y_1, Y_2 and Y_3 to distinguish four groups based on many objects of 144 districts, many groups of 5, and the large of lambda 0.006944. The average of discriminant coefficients for each variable in each vertex discriminant function can be seen in Table 4.

Table 4 Average of discriminant coefficient of each variable to vertex discriminant function

Discriminant coefficients	Y_1	Y_2	Discriminant coefficients	Y_1	Y_2
	Mean	Mean		Mean	Mean
Intercept $(\bar{b})$	0,142	0,146	$\bar{a}_{14}$	0,005	-0,060
$\bar{a}_1$	0,020	-0,063	$\bar{a}_{15}$	0,019	0,009
$\bar{a}_2$	0,129	0,035	$\bar{a}_{16}$	0,041	0,048
$\bar{a}_3$	0,008	0,113	$\bar{a}_{17}$	-0,003	-0,013
$\bar{a}_4$	0,090	0,001	$\bar{a}_{18}$	-0,021	0,026
$\bar{a}_5$	0,047	-0,011	$\bar{a}_{19}$	0,046	-0,041
$\bar{a}_6$	-0,021	0,114	$\bar{a}_{20}$	0,040	-0,037
$\bar{a}_7$	0,050	-0,068	$\bar{a}_{21}$	-0,017	0,048
$\bar{a}_8$	-0,021	-0,012	$\bar{a}_{22}$	-0,006	-0,085
$\bar{a}_9$	0,030	-0,040	$\bar{a}_{23}$	0,031	0,057
$\bar{a}_{10}$	-0,004	-0,017	$\bar{a}_{24}$	0,013	0,004
$\bar{a}_{11}$	-0,023	0,013	$\bar{a}_{25}$	0,027	-0,004
$\bar{a}_{12}$	0,078	-0,050	$\bar{a}_{26}$	-0,002	0,029
$\bar{a}_{13}$	0,014	0,011	$\bar{a}_{27}$	-0,008	0,004

By determining the simultaneous confidence intervals of each discriminant coefficient at the vertex function, can be seen the significant variables. The confidence interval for the discriminant coefficients which contains a value of zero indicates that the variable is not significant. The simultaneous confidence interval 95% for each coefficient discriminant variables at each vertex discriminant function can be seen in Figure 1.



(a) (b)

Figure 1 (a) The simultaneous confidence interval for the discriminant coefficients of Y_1
(b) The simultaneous confidence interval for the discriminant coefficients of Y_2

From Figure 1, it is seen that in the discriminant function of vertex Y_1 the percentage indicator of household with electricity users (X_{10}), the number of health facilities per 1000 population (X_{14}), financial capacity (X_{17}), the percentage of vulnerable villages of landslides (X_{22}) and the percentage of critical land village (X_{26}) are not significant. In the discriminant function of vertex Y_2 the indicator of the average length of school (X_4), the literacy rate (X_5), the percentage of vulnerable village of other disaster (X_{24}), the percentage of villages in protected areas (X_{25}) and the percentage village of conflict in recent year (X_{27}) are not significant. However, there is no a variable that is not significant at the third vertex functions simultaneously. Thus, it can be concluded that all variables give significant effect. In other words, all variables can be included in the formation of discriminant function.

Based on 27 variables obtained, the form of the first discriminant function (Y_1) and the second discriminant function (Y_2) as follows:

$$\begin{aligned} Y_1 = {} & 0{,}184 + 0{,}057X_1 + 0{,}176X_2 - 0{,}011X_3 + 0{,}096X_4 + 0{,}05X_5 + 0{,}0273 + \\ & 0{,}033X_7 - 0{,}055X_8 + 0{,}047X_9 - 0{,}016X_{10} - 0{,}025X_{11} + 0{,}115X_{12} + 0{,}037X_{13} - \\ & 0{,}059X_{14} + 0{,}059X_{15} + 0{,}034X_{16} - 0{,}002X_{17} - 0{,}051X_{18} + 0{,}07X_{19} + 0{,}055X_{20} - \\ & 0{,}033X_{21} - 0{,}006X_{22} + 0{,}028X_{23} + 0{,}003X_{24} + 0{,}042X_{25} + 0{,}05X_{26} - 0{,}003X_{27} \end{aligned}$$

$$\begin{aligned} Y_2 = {} & 0{,}108 - 0{,}086X_1 + 0{,}019X_2 + 0{,}126X_3 - 0{,}04X_4 + 0{,}026X_5 + 0{,}052X_6 - \\ & 0{,}072X_7 - 0{,}011X_8 - 0{,}06X_9 - 0{,}019X_{10} + 0{,}022X_{11} - 0{,}052X_{12} - 0{,}004X_{13} - \\ & 0{,}035X_{14} - 0{,}007X_{15} + 0{,}054X_{16} - 0{,}013X_{17} + 0{,}02X_{18} - 0{,}057X_{19} - 0{,}034X_{20} + \\ & 0{,}064X_{21} - 0{,}095X_{22} + 0{,}055X_{23} + 0{,}0001X_{24} + 0{,}025X_{25} + 0{,}031X_{26} + 0{,}011X_{27} \end{aligned}$$

In the function of Y_1 coefficient discriminant of the percentage indicator of poor citizen (X_1) is 0.057, while the function of Y_2 is -0.086. Positive and negative sign indicate the direction of the effect of the indicator. The percentage indicator of poor citizen (X_1) is an indicator that measures the badness. The greater value of the discriminant coefficient (positive) will encourage district status to be worse and the smaller value of the discriminant coefficient (negative) will encourage district status to be better. And also for the indicator that measures the goodness. For example, the indicator of life expectancy (X_3) in the function of Y_1 discriminant coefficient value is -0.011 and in the function of Y_2 is 0.126. The indicator of life expectancy is an indicator that measures the goodness. The greater value of the discriminant coefficient (positive) will push district status to be better and the smaller value of the discriminant coefficient (negative) will push district status to be worse. Similarly, for the indicator that measures the goodness and the indicator that measures the other badness.

The goodness of VDA discriminant function method above can be seen from the classification of accuracy of each group as seen in Table 6.

Table 6 Goodness of VDA discriminant function method for 144 training data

		Model Classification			
	Group	1	2	3	Many Objects
The Real Classification	1	100	2	1	103
	2	2	17	1	20
	3	2	0	19	21
	Many Objects	104	19	21	144

Based on Table 6, it can be seen that VDA discriminant function method is able to classify 136 objects (94,44%) accurately and many objects are misclassified about 8 objects (5,56%). Based on the attachments, it can be seen that the error classification of (value APER) VDA method about 0.055556. It shows that the accuracy of VDA Method in classifying recisely objects in the case of data training about 0.944444.

Furthermore, the evaluation of VDA method was done by viewing the great error classification (APER value) by using 18 data testing. Such as for the first object of testing data is South Solok district with the initial classification is less underdeveloped districts. The value for each variable and the value of standardization can be seen in Table 7.

Table 7 Values of variables for the first object of data testing (South Solok)

Variables	Data	Standardized data
X_1	8,12	-0,90
X_2	623,15	0,43
X_3	64,94	-1,47
X_4	8,17	1,03
X_5	97,72	1,04
X_6	29	-0,79
X_7	6	-1,05
X_8	4	-0,67
X_9	0	-0,45
X_{10}	92,31	0,93
X_{11}	5,91	1,37
X_{12}	82,05	1,51
X_{13}	5	-0,46
X_{14}	0,1	-0,46
X_{15}	0,02	-0,71
X_{16}	0,04	-0,83
X_{17}	221542,04	-0,21
X_{18}	30,53	-0,61
X_{19}	23,27	0,43
X_{20}	5,90	-0,47
X_{21}	25,64	2,21
X_{22}	7,69	-0,15
Variables	Data	Standardized data
X_{23}	14,00	-0,33
X_{24}	51,28	1,23
X_{25}	0,00	-0,69
X_{26}	0,00	-1,20
X_{27}	10,26	0,39

The value of each variable is substituted into 2 vertex discriminant function to obtain the value $Y_1 = 0{,}234$, and $Y_2 = 0{,}043$. Then calculated the shortest distance between the third value of vertex discriminant function toward the node to determine the assumption of classification of South Solok district. The value of the node to 3 groups and the distances between objects with the node can be seen in Table 8.

Table 8 The node and the distance between objects with the node

Group		$\boldsymbol{v}_j$	$\boldsymbol{Y}_j$	$\left\|\boldsymbol{v}_j - \widehat{\boldsymbol{A}}_{x_i} - \widehat{\boldsymbol{b}}\right\|$
1	$\boldsymbol{v}_1$	$\begin{bmatrix}0{,}707\\0{,}707\end{bmatrix}$	$\begin{bmatrix}0{,}234\\0{,}043\end{bmatrix}$	0,816
2	$\boldsymbol{v}_2$	$\begin{bmatrix}0{,}258\\-0{,}965\end{bmatrix}$	$\begin{bmatrix}0{,}234\\0{,}043\end{bmatrix}$	1,008
3	$\boldsymbol{v}_3$	$\begin{bmatrix}-0{,}965\\0{,}258\end{bmatrix}$	$\begin{bmatrix}0{,}234\\0{,}043\end{bmatrix}$	1,218

Based on Table 8 by looking at the smallest distance between the object and the node, South Solok district is classified into group less underdeveloped districts where the assumption of classification is the same as the initial classification. The results of classification for 18 data testing can be seen in Table 9.

Table 9 The classification result of VDA discriminant function method for 18 data testing

	Model Classification				
	Group	1	2	3	Many Objects
The real classification	1	10	1	1	12
	2	1	2	0	3
	3	1	1	1	3
	Many Objects	12	4	2	18

Table 8 describes that on data testing 13 objects (72,22%) appropriately classified to the third VDA discriminant function method and 5 objects (27,78%) misclassified. Thus, we can conclude that VDA discriminant function method is good method in distinguishing groups.

However, in the same way if the variables are not significant not included in the formation of vertex discriminant function, the goodness of classification result in training data is 86.11% and the goodness of classification in testing data is 72.22%. The goodness in classification on training data decreases when the variables are not significant eliminated. However, in testing data the goodness of classification is the same. It shows that although a variable is not significant in one vertex function, the variable still can be included in the formation of discriminant function because the vertex function of the variables is still significant.

CONCLUSION AND SUGGESTION

Based on the research result of the methods of vertex discriminant multicategory analysis with bootstrap application, it can be concluded as follows:

1. 27 variables or indicators used by the ministry in determining the status of underdeveloped district, provide the real influence in the formation of vertex discriminant function.
2. The accuracy of the classification of vertex discriminant function that is formed about 94,44 percent for training data and 72,22 percent for testing data. In the other words, to determine the newest underdeveloped district status, Ministry of Rural Development can use 27 available indicators.

Suggestion for further research:

In this research, it is known that the VDA method has good ability in classification. VDA method can be applied in various fields. In the case of multivariate classification that involve many variables need to be done the significance test to reduce variables so that the resulting discriminant function more efficiently.

REFERENCE

Central Bureau of Statistics. (2013). *Statistik Keuangan Kabupaten/Kota Tahun 2012-2013*. Jakarta, Indonesia: Author.

Central Bureau of Statistics. (2011). *Statistik Potensi Desa Provinsi 2011*. Jakarta, Indonesia: Author.

Central Bureau of Statistics. (2014). *Statistik Potensi Desa Provinsi 2014*. Jakarta, Indonesia: Author.

Davison, A. C. And Hinkley, D. V. (1997). *Bootstrap Methods and their Applications.* Cambridge University Press.

Efron, B. And Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* New York: Chapman & Hall.

Johnson RA, Wichern DW. (2007). *Applied Multivariate Statistical Analysis (*6th ed). New Jersey (US): Pearson Prentice Hall.

Ministry of Rural Development. (2010). *Rencana Strategis Tahun 2010-2014*. Jakarta, Indonesia: Author.

Nurmaleni. (2015). *Perbandingan Metode Klasifikasi antara Analisis Diskriminan verteks dan Diskriminan Fisher*. (Masters theses). Available from Bogor Agricultural University repository.

Lange K, Wu TT. (2008). An MM Algoritm For Multicategory Vertex Discriminant Analysis. *J Comput Graph Stat.* 17(3): 527-544.

Lange K, Wu TT. (2010). Multicategory Vertex Discriminant Analysis For High-Dimensional Data. *The Annals of Applied Statistics.* 4(4): 1698-1721.

Rencher AC. 2002. *Methods of Multivariate Analysis* (second ed). New York (US): Wiley.